Ondertitelen UvN-video's

Achtergrond

Op verzoek van de Nederlandse Taalunie (NTU) zijn meer dan 50 video's van de Universiteit van Nederland (UvN) ondertiteld en op woordniveau doorzoekbaar gemaakt. Bij de Universiteit van Nederland delen inspirerende wetenschappers van Nederlandse Universiteiten de spannendste inzichten uit hun vakgebied met een jong en breed geïnteresseerd publiek door wekelijks een prikkelende vraag in een college van ong. 15 minuten te beantwoorden.

Dit goed toegankelijke materiaal is niet alleen boeiend om de gepresenteerde inhoud, maar ook zeer geschikt om gebruikt te worden voor het leren van Nederlands als tweede taal. Docenten Nederlands in het buitenland hebben de NTU gevraagd of er ook ondertitels van de video's beschikbaar waren om op die manier het materiaal voor het geven van lessen Nederlands geschikt te maken.
Dat bleek helaas niet het geval en dus moesten alle ondertitelingen van scratch gemaakt worden.

Het ondertitelen werd gedaan door eerst alle audio door de Automatische Spraakherkenner (ASR) te halen. Hiermee werd een ruwe, doorgaans goede transcriptie verkregen, echter zonder interpunctie of hoofdletters. De transcripties werden vervolgens handmatig door studenten van de Radboud Universiteit gecontroleerd en gecorrigeerd. Ten slotte werden de uiteindelijke transcripties opnieuw door de Spraakherkenner verwerkt maar nu in de zgn FA-mode: Forced Alignment. Hierbij blijft de tekst onveranderd maar wordt van ieder opgeschreven woord de juiste begin- en eindtijd berekend, hetgeen nodig is als uitgangspunt voor een ondertiteling.

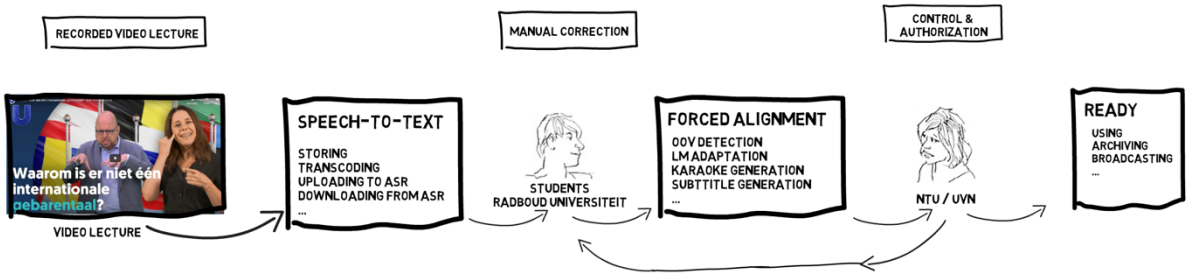
Het resultaat na de Forced Alignment is per video een rij woorden die volgens de menselijke controleurs daadwerkelijk werden uitgesproken, incl. de begin- en eindtijd van elk woord. Van deze rij woorden werden vervolgens de ondertitelingsbestanden gemaakt. De gekozen ondertitelingsbestanden (SRT & VTT) kennen een begin- en eindtijd op ondertitelingsregelniveau.

De ondertitelingen werden gemaakt volgens de door de Nederlandse Auteursbond gepropageerde richtlijnen waarbij er niet meer dan 2 regels op het scherm getoond werden, iedere regel met meer dan 45 karakters in tweeën werd gedeeld en het totaal aantal karakters op het scherm onder de 100 bleef.

Materiaalselectie

De selectie van de video's is een gecombineerde inspanning geweest van zowel docenten Nederlands in het buitenland als mensen van de UvN. De docenten Nederlands willen het materiaal voor educatieve doeleinden gebruiken. De UvN wil inzicht krijgen in de mate waarin ondertiteling en doorzoekbaarheid van het materiaal, voor een betere ontsluiting en hergebruik van de video's kan zorgen.

Werkproces



Figuur 1: Workflow voor het ondertitelen van de videolezingen van de Universiteit van Nederland

De geselecteerde video's werden in een periode van een paar weken door een medewerker van de UvN op de Surfdrive van het project gezet. Dit duurde langer dan gehoopt omdat het Coronavirus roet in het eten gooide:

niet alle bestanden stonden online en het kantoor was gesloten. Uiteindelijk is het wel gelukt alle gewenste video's te krijgen. Tegenvallend was wel dat het formaat en de naamgeving van de video's nogal wisselend was en dat de metadata (bv. wie spreekt er?) volkomen afwezig was (veel video's heetten bv **opname_uvn** oid).

HARMONISATIE

De eerste stap was daarom het harmoniseren en uniformeren van de bestandsnamen. Gekozen werd de (lange) titel van de video te gebruiken voor alle bestanden die met die video te maken hebben. Als voorbeeld kan de video van Marc van Oostendorp genomen worden. De oorspronkelijke naam luidde: "**Waarom is het raar om met een Gooise R te praten?**" en de video-file heette "**college_MvO_1.mov**".

Figuur 2: voorbeeld van een directory met daarin alle verschillende bestanden. Ze heten allemaal naar de directory waarin ze horen te staan zodat duidelijk is waarbij ze horen. In het voorbeeld hierboven staat een volledige set van op te leveren bestanden.

Van de officiële naamgeving werden de leestekens (? & !) verwijderd, de woorden in een standaard ASCII manier herschreven (überhaupt -> uberhaupt) en de spaties werden vervangen door underscores. De naam van de directorie waarin alles van deze video verzameld werd luidde daarna:
Waarom_is_het_raar_om_met_een_Gooise_R_te_praten en alle files in deze directory heette vervolgens ook zo (zie figuur 2).

Omdat de oorspronkelijke videobestanden in verschillende formaten stonden (bv. MP4, MOV, MPEG, OGG) werden **alle** video's opnieuw getranscodeerd naar het moderne MP4 (H.265) formaat.

De video's in dit project waren bedoeld voor handmatige controle door thuiswerkende studenten en moesten daarom niet te groot worden om ze snel te kunnen down- en uploaden. Besloten werd ze allemaal te downschalen naar 800 x 600 pixels met 4000Kb/s. De gemiddelde grootte van de videobestanden daalde van meer dan 1 GB naar 60MB waarmee de omvang goed hanteerbaar werd.

De ASR (automatische spraakherkenner) gebruikt audio-bestanden in het formaat 16kHz, 16-bit sampling en mono. Van alle video's werd daarom het audiospoor apart opgeslagen als 16-16-1 WAV-file.

Het resultaat van deze harmonisatie was een verzameling van directories met daarin de getranscodeerde MP4-video-file en de geripte WAV-audio-file en een uniforme naamgeving van alle bestanden en folders.

Spraakherkenning

Alle audiobestanden werden in batches van 10-files per keer door de ASR-server gehaald. Het resultaat van iedere herkenning was een CTM-file: een met CSV vergelijkbare tekstfile met de volgende kolommen:

ID (naam), nr (altijd 1), begintijd van het herkende woord in msec, duur van het herkende woord in msec, het herkende woord, en de betrouwbaarheid van de herkenning (0 <>1 waarbij 1 staat voor zeer betrouwbaar). In figuur 3 staat een screenshot van een CTM-file.

```
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 2.76 0.12 de 0.96
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 2.88 0.57 opwarming 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 3.45 0.15 van 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 3.6 0.09 de 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 3.69 0.39 aarde 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 4.08 0.24 ook 0.60
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 4.32 0.69 tegengaan 0.95
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 5.1 0.24 door 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 5.34 0.51 zonlicht 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 5.85 0.09 te 0.99
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 5.94 0.66 weerkaatsen 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 8.35 0.20 dit 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 8.55 0.18 is 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 8.73 0.15 de 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 8.91 0.72 universiteit 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 9.63 0.15 van 1.00
Kun_je_de_opwarming_van_de_aarde_ook_tegengaan_door_zonlicht_te_weerkaatsen 1 9.78 0.63 Nederland 0.97
```

Figuur 3: voorbeeld van de eerste 17 herkende woorden in het CTM-formaat.

Deze CTM-file werd vervolgens ingelezen in ASR-corrector; een speciaal voor dit project geschreven softwareprogramma waarmee de spraakherkenning snel en relatief makkelijk gecorrigeerd kan worden.
De uiteindelijke werkwijze was als volgt: ASR-corrector las steeds 100 herkende woorden in en maakte daar een chunk van. Een chunck bestaat uit een deel van de opname met een begintijd (*begintijd van het eerste woord*) en eindtijd (*eindtijd van het 100ste woord*).

Correctie

Deze chuncks van 100 woorden werden gecontroleerd en waar nodig verbeterd. Ook werd voor iedere chunck de hoofdspreker aangegeven en werden de sprekerswisselingen in de chunck-tekst zelf aangegeven.

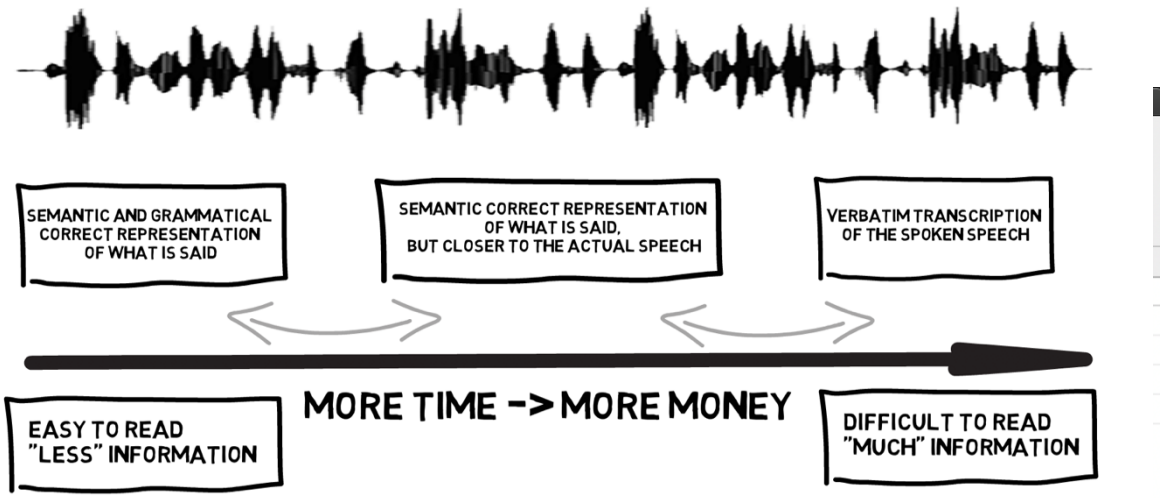
De transcribenten verbeterden met de hand de verkeerd herkende woorden en zochten soms bij vreemde namen, de officiële schrijfwijze op. Doordat het herkenningsresultaat van de spraakherkenner geen woorden met een hoofdletter kent, moesten alle woorden die altijd met een hoofdletter geschreven worden, allemaal herschreven worden (bv nederland -> Nederland). Op verzoek van de studenten werd dit geautomatiseerd zodat in een keer alle voorkomens van “amsterdam” als “Amsterdam” werden herschreven. Het transcriptieprotocol is als apart bestand aan deze documentatie toegevoegd.

Zin-creatie

Het resultaat van de spraakherkenning is een rij opeenvolgende woorden zonder leestekens. Een van de opdrachten die de transcribenten meekregen was het “maken van zinnen voor zover mogelijk”. Dit is en blijft een lastig proces omdat mensen nu eenmaal niet in zinnen spreken en vrij veel sprekers de neiging hadden eindeloze en, en, en “zinnen” te maken of om zichzelf te verbeteren “*of nou nee, laat ik het zo zeggen*”.
Ook bleef het steeds een afweging hoe letterlijk de tekst getranscribeerd moet worden. Wordt een herhaling van 3x “de” ook zo opgeschreven of niet? Sommige sprekers waren grammaticaal vrij correct maar andere juist niet. En een al te letterlijke transcriptie kan dan de leesbaarheid sterk verminderen.

Om het proces van “zinnenmaken” te vereenvoudigen, werd in ASR-corrector een routine gemaakt waarmee met de rechtermuis een zin gemaakt kon worden (laatste woord eindigt met een punt, volgende woord begint met een hoofdletter). Vraagtekens en komma’s moesten handmatig worden toegevoegd.

Het werd aan de studenten overgelaten de zin-creatie zo goed mogelijk te doen. De zinsopbouw werd later steekproefsgewijs gecontroleerd en waar nodig aangepast door de tweede corrector.



Figuur 4: Schematisch overzicht van de verschillende transcriptiemogelijkheden. In dit UvN-project werd gekozen voor een behoorlijk verbatim transcriptie. Ten eerste omdat daarmee het materiaal geschikter werd voor onderwijsdoeleinde en ten tweede omdat de meeste sprekers hun praatjes goed hadden voorbereid en dus grammaticaal netjes spraken. Een bijna verbatim transcriptie deed daarom weinig af aan de leesbaarheid.

TWEEDE CORRECTIE

Wanneer de studenten een transcriptie gecorrigeerd hadden, werd het bestand weer op de gezamenlijke SurfDrive geplaatst maar dan in een andere directory (ASR_gecontroleerd). Een tweede corrector controleerde en corrigeerde steekproefsgewijs de resultaten, en bracht de student daarvan op de hoogte (*volgende keer opletten op…*). Dit om de studenten te helpen en om een zeker mate van uniformiteit te bereiken.

Nadat de tweede corrector tevreden was werden de bestanden verplaatst naar opnieuw een andere directory: ASR_klaar (zie ook schema in figuur 1).

```
[StartTime: 4080ms]
[Speaker: B]
Hoe kwam er een einde aan de Gouden Eeuw? Dit is de Universiteit van Nederland. Twintig augustus 1672, het
gonst van de geruchten in Den Haag. Boeren uit het Westland zouden op weg zijn naar de stad. De schutterij
wordt onder de wapenen geroepen om dat te verhinderen. Maar er zijn ook allerlei schimmige figuren in Den
Haag actief. Gedurende de voorgaande dagen en weken hebben die vergaderd op allerlei locaties. Om... Ja
wat precies? Daarvan zijn natuurlijk geen documenten overgeleverd maar uit wat er daarna gebeurt is kunnen
we dat wel ongeveer reconstrueren. Één van die mensen is Cornelis Tromp, een vlootvoogd,
[EndTime: 57550ms]
[StartTime: 57580ms]
[Speaker: A]
een iemand die leiding heeft gegeven aan de marine. Successen heeft gevierd maar die intussen in ongenade
is gevallen. Op één van die vergadering is ook Willem Derde himself gesignaleerd: op dat moment de
opperbevelhebber van het leger, Prins van Oranje, iemand die in de voorafgaande maanden een enorme
uitbreiding van zijn bevoegdheden in zijn macht heeft meegemaakt. In dat klimaat beweegt zich halverwege
de ochtend Johan de Witt van zijn huis naar de gevangenpoort. Johan de Witt is de voorafgaande ruim
twintig jaar zeg maar, de politieke leider geweest van de Republiek; de Mark Rutte van
[EndTime: 104250ms]
```

Figuur 5: Screenshot van de tekstfile die door ASR-corrector gemaakt wordt nadat de correcties klaar zijn. De file kent een aantal metadata (tijd, spreker) die allemaal tussen [] staan zodat ze makkelijk geparsed kunnen worden.

Nadat de correcties gedaan waren, werd het resultaat geëxporteerd naar een speciale tekst-file (zie figuur 5) waarbij metadata bewaard werd tussen []. Deze speciale tekst-files dienen als input voor de volgende stap: Forced Alignment.

Forced Alignment

Na de hierboven beschreven stappen was er een correcte transcriptie per chunk (van oorspronkelijk 100 herkende woorden met gemiddeld tussen de 450 en 550 karakters). De tijdsduur van elke chunk was veel te groot voor adequate ondertiteling en bovendien wilde we graag een bestand opleveren waarbij van **ieder woord** precies bekend was wanneer het werd uitgesproken.

De gecorrigeerde teksten en de audio werden daarom onderworpen aan een Forced Alignment (FA). De input waren de speciale txt-files (Figuur 5) en de output van de FA-service waren opnieuw CTM-files maar nu met uitsluitend de juiste woorden, leestekens en hun bijbehorende start- en eindtijd.

Figuur 6: Schematische weergave van Forced Alignment. Een bestaande tekst wordt zo goed mogelijk op tijd gezet binnen de aangegeven grenzen.

Eindresultaat

De laatste stap die gemaakt moest worden is de export van de FA-CTM in de verschillende gewenste formaten: HTML, SRT en VTT.

Dit werd gedaan met FA-convertor: een voor dit project geschreven programma dat de tekstfile (figuur 5) en de FA-CTM-resultaten combineerde.

Het uiteindelijke resultaat bestaat per video uit de volgende 7 bestanden:

EXTENSIE	BESCHRIJVING
MP4	De video-file zoals getranscodeerd na levering door de UvN. De mp4 werd gedurende het gehele proces gezien als HET oorspronkelijke bronmateriaal.
WAV	De audiofile van de opnamen. De wav werd altijd uit de getranscodeerde mp4 gehaald
TXT	De tekst-file (zoals hierboven getoond) incl wat metadata over de start en eindtijd van de chunck en de sprekers die in die chuck aan het woord komen. De txt-file is waarschijnlijk het best voor alleen lezen.
HTML	De HTML of zogenaamde Karaoke file is een combinatie van een tekst-file en een kort java-scriptje dat, wanneer er op een woord gedrukt wordt, vanaf dat woord de file gaat afspelen. Ieder woord dat woord “voorgelezen” wordt gehighlight. Het is de ideale manier om lezen en luisteren te combineren.
XML	Het bronbestand voor de verwerking en correctie. De CTM (resultaat van de spraakherkenning) werd door ASR-corrector ingelezen en omgezet in een bruikbaarder xml-formaat. De correcties van de studenten werden ook in de xml gestopt. De uiteindelijke xml-files bevatten daarom zowel de correcte teksten als de oorspronkelijke teksten zoals gegeneerd door de spraakherkenner.
SRT	De uiteindelijke ondertiteling in het SubRipType format. SRT is de default standaard en kan door alle normale video-software gebruikt worden
VTT	Een sterk op SRT lijkend formaat dat gebruikt wordt voor de ondertiteling van video op het Internet.

Analyse

De aanwezigheid van de CTM-files waarin van elk woord de begin- en eindtijd bekend is, biedt de mogelijkheid om makkelijk (eenvoudige) analyses te doen. Zo bleek men, op een paar uitzonderingen na, redelijk constant te spreken met gemiddeld 1.56 woorden/sec (≈93 woorden/minuut, hetgeen redelijk langzaam is). In een gemiddelde lezing worden er 130 woorden/minuut gesproken (zie hier). Ook de spreektijd bleek redelijk constant.

	#Woorden	*Duur in sec*
Totaal	151259	97078.228
Gemiddeld	2701	1733.540